A summary of the February 21 spam retrospective

Leigh Honeywell posted an open invitation for Fediverse admins to meet and discuss the spam crisis that happened over the weekend of February 15, 2024. A bunch of us met on Zoom on Feb 21. Here's what we talked about, as summarized by Jason McIntosh.

What happened and when

The first reports of spam from Mastodon server admins started showing up on the afternoon (UTC) of Tuesday, February 13. Others started receiving spam reports on Thursday the 15th, and the flow stayed pretty thick through the whole long weekend.

Almost every piece of spam used one of a fixed handful of images as its content, sometimes with some predictable text or hashtags attached. Most of the spamming instances seemed to be "derelict" Mastodon servers without anyone actively moderating them.

The rate of spam lessened dramatically after Tuesday, February 20. There were some "second wave" spams that followed similar patterns in the days afterwards.

Why this happened

Two kids had a fight over in Discord-land. One got the idea to attack the other by launching a script that would spam Misskey servers with messages appearing to come from the other kid's Discord server, a new spin on an old-school joe job. The script worked well enough that its output spilled into Mastodon over their shared ActivityPub protocol.

- The code used in the attack, hosted on GitHub.
- A thorough investigation into the context and motivations behind the incident.

How admins responded

Admins applied a mix of shared blocklists, purpose-made automation, and manual labor to combat the tide, with generally positive results. The predictable content of the spam and patterns of the fake accounts made this feasible.

Admins also shut off open registration, in order to prevent their own servers from being used to send more spam.

Many admins made do with Mastodon's native administrative tools, which allowed the import of community-made blocklists. However, admins also felt hampered by the lack of common Fedi tools and processes specific to spam prevention. The work was painfully manual for everyone. Admins had to find their own ways connect with the larger community and form ad-hoc anti-spam strategies mid-crisis.

Many admins felt very hesitant to mass-block well-intentioned servers that had been turned by malefactors into spam-launchers, as there were no community standards about how to respond to this situation.

Impact

Some server admins had to slog through dozens or hundreds of user reports during the peak of the crisis. Others barely saw anything.

Admins who did have to deal with the deluge spent at least a few hours over the weekend cleaning it up, often in brief shifts spread throughout the three-day period. Those who also wrote custom tools to deal with the spam ended up putting in much longer hours.

How we feel in retrospect

All things considered, this could have been so much worse. The spam content was relatively inoffensive. While the amount of spam was both annoying to behold and worrying in its implication, it was not so great in volume that it literally DDoSed any servers. The spams' static content also made its identification and cleanup easier.

All of which is say: a future attack made using the same methods could, in fact, be so much worse.

Official response

The Mastodon software project has been responding to the crisis quickly with new and proposed features appearing on the project's GitHub repository. As of February 21, these included the following:

- Change registrations to be disabled by default for new servers
- Automatically switch from open to approved registrations in absence of moderators
- Add support for server-specific emergency rules

Community response

- IFTAS is on the cusp of FediCheck 1.0, a domain-block manager.
- Various folks are building an ISAC for the social web.
- Others are working on a system for sharing fediverse-wide moderation advisories.